1-1-2000

# Volume 11, Number 1 & 2 - CogSci News (Spring 2000)

Lehigh University Cognitive Science Program

# CogSci News

**Cognitive Science Program, Lehigh University, Bethlehem, PA.**

*Volume 11, Number 1 & 2*
*Fall 2000*



### Editorial Policy

This newsletter is published once or twice yearly by the Cognitive Science Program at Lehigh University. Its purpose is to provide a forum for discussing issues and developments in cognitive science and to report the activities of Lehigh's Program.

The newsletter is distributed free of charge in the United States and Canada to academic programs and individuals interested in cognitive science. To be added to the mailing list, simply fill out the form on the following Web page: http://www.lehigh.edu/~incog/subscription.

The Editorial Staff welcomes readers' comments and *solicits materials* dealing with cognitive science. We are especially pleased to consider program descriptions, short essays, brief descriptions of scholarship and research in progress, book reviews, and original art work (e.g., cartoons, line-drawings, computer-generated images).

Address all submissions and comments to the editor: Padraig G. O'Seaghdha, CogSci News, Lehigh University, 17 Memorial Drive East, Bethlehem, PA 18015. Or send electronic mail to: pat.oseaghdha@lehigh.edu.

## EDITORIAL: Fringe Benefits

*CogSci News* is back after an unscheduled absence of three years. We have certainly had plenty of time to take stock and reassess our mission. This editorial redefines the niche of *CogSci News* based on a brief assessment of the overall status of the field.

Cognitive Science as a discipline appears to be both making substantial, even surprising progress, and to be in something of a rut. The term Cognitive Science has clearly established purchase on most things cognitive and Cognitive Science programs are expanding worldwide. *Trends in Cognitive Sciences* is a lovely, accessible digest of new research. These are just two indications that the mainstream is thriving.

On the other hand, Cognitive Science still appears to be stuck on the horns or multiple prongs of disciplinary/interdisciplinary dilemmas. Much of the progress appears to be in specific disciplines which stubbornly refuse the role of tributaries to Cognitive Science itself. This is perhaps most unfortunately obvious in the Cognitive Science Society Conference which for the most part functions as a forum for two primary disciplines, Artificial Intelligence and Cognitive Psychology. Lots of good work in these areas is presented, but somehow the opportunity for creating a true interdisciplinary nexus for Cognitive Science has not been seized. The obvious downside of this is that many of us choose to attend more focused disciplinary conferences instead of the one that should represent the most adventurous, cross cutting, projects and ideas.

A suggestion to the organizers of the next Cognitive Science Society conference in Edinburgh: The Edinburgh Arts Festival has become justly world famous because it provides an inclusive, open organization that wholeheartedly welcomes the Fringe while also promoting more established mainstream acts. Cognitive Science is now surely mature enough to be able to indulge its wilder, more rambunctious side, while also accommodating a more stolid and dignified core. This is not to advocate *cognitive psoriasis* (the dreaded flake disease), but rather perpetuation of the kind of creative ferment that defined the beginnings of cognitive science. For Cognitive Science to prosper, it must remain true to its original boundary-oblivious impulse.

Locally, Cognitive Science at Lehigh is also in a state of mixed perplexity and hope. In some respects, the program has been marginalized as resources in the last decade have been concentrated in Departments. However, Lehigh has just announced a major thrust to reinvigorate academic programs generally, and we are optimistic that interdisciplinary programs like Cognitive Science will be substantial beneficiaries of this development. A prospective University-wide initiative in Information Science and Technology is especially promising for Cognitive Science.

Looking to the future of this newsletter and its raison d'etre, *CogSci News* has always served a dual role that we now more explicitly define.

First, we provide a forum for discussion and dissemination of programmatic and curricular matters, the latter exemplified in this issue by our twin position papers on teaching Introductory Cognitive Science. We invite submissions of such pedagogical pieces, as well as new program descriptions, focusing especially on innovations and their underlying rationale. The easy accessibility of program details on the internet makes mere description of program nitty gritty (except through posting of links) redundant, and

(continued on page 2)

## Editorial (cont.)

frees authors to focus on motivating ideas instead. **Submissions on these lines are hereby solicited.**

Second, we provide an outlet for exploratory, Fringe Cognitive Science, in the positive sense pointed to above. Cognitive Science is reaching its tentacles into new regions that are rarely represented in our conferences and journals. For example, Lehigh's 1999 Cognitive Science Keynote Speaker, Dan Gilbert, introduced us to the idea of Social Cognitive Neuroscience, a subfield label that was surprising just two years ago, but may be already failing to raise eyebrows. Our 2000 Keynote Speaker, Mark Turner, is heavily involved in another relatively unknown but burgeoning area, Cognitive Science and Literature (see for example, http:// cogweb.english.ucsb.edu/). These are just two examples of topics that *CogSci News* may serve the function of disseminating to the wider Cognitive Science community in future issues. **Again, we invite submissions and inquiries.**

The ascendancy of the internet also raises a practical question of how best to distribute *CogSci News* in future. We are already simulcasting *CogSci News* on the web as well as on paper. The next issue of *CogSci News* will likely be the last to be distributed widely in hard copy. After that we will announce each new electronic issue by email and other means.

*CogSci News* is alive and kicking. However, it needs the food of articles to survive and thrive. Please send your contribution soon.

— P.O'S.

# Teaching Introductory Cognitive Science at Lehigh University: Two Approaches

Ideally, given the interdisciplinary nature of cognitive science, any introduction at the undergraduate level should be team-taught. Such was the practice at Lehigh for several years following the inception of our program (see earlier pieces by Malt & Melchert, 1988, and Kay, 1992, in *CogSci News*). As increased budgetary consciousness came to weigh against such collaborations, however, it has become usual to assign the course to a single instructor. Each instructor must now contend not only with his or her own disciplinary limitations, but with the task of presenting a unified survey of cognitive science without other-disciplinary colleagues to lean on. Most recently, this Herculean task has been assumed by Mark Bickhard and Alex Levine who here report on their distinct *historical* and *thematic* approaches to the problem of being all of Cognitive Science to novices.

---

### A Historical Approach to Teaching Introductory Cognitive Science

Mark H. Bickhard
Lehigh University
( mark.bickhard@lehigh.edu )
http://www.lehigh.edu/~mhb0/ mhb0.html

Introductory Cognitive Science at Lehigh generally has a broad mix of students, both with respect to their majors and their years in school. But the course is intended to attract first year students in particular, and this has produced a problem in teaching. Specifically, available textbooks for Cognitive Science seem to be inappropriate for one or both of two reasons: 1) they are not genuinely interdisciplinary cognitive science texts, but, instead, focus primarily on one of the affiliated disciplines, usually artificial intelligence, or 2) they are too advanced for most first year college students.

I have found two books that offer an interdisciplinary introduction: Stillings, Weisler and Chase's (1995) *Cognitive Science: An Introduction* and Ó Nualláin's (1995) *The Search for Mind.* Stillings et al. provides a powerful overview that is truly interdisciplinary, but it moves too strongly into advanced discussions that first year students find unacceptably difficult. It is also difficult to isolate the advanced passages and to focus on the more introductory discussions because they are thoroughly mixed. Stillings et al is a book with a grand sweep that would fit extremely well in an introductory course for more advanced students.

Ó Nualláin is a shorter book, but it too moves too fast and assumes too much for use in an introductory course. Ó Nualláin is also a polemical book, and so would serve well in an advanced course, both for filling out narrower horizons and for stimulating discussion: there is something to outrage everyone, and it will surely keep interesting controversy going (Bickhard, 1997).

In response to this difficulty in finding a single interdisciplinary text, I have adopted a historical approach to introducing Cognitive Science, with several texts selected to fill out this history, and no privileged core text. The course begins with Gardner (1985). Gardner is powerfully interdisciplinary, and explicitly historical in its approach. He slights, in my judgment, the influence of Piaget early in the history of cognitive science, and, like all authors, has his own definite axes to grind, but the book is overall a good beginning. Students like it, and it provides a broad background for the contributions and relevance of all of the major disciplines associated with cognitive science. There are two related problems with Gardner: 1) it only follows the history up to the middle 1980s, and, correspondingly, 2) in light of more recent developments, some of the discussions of the issues at the time of publication now have a slightly dated air. Nevertheless, it has worked well over several years of using this historical approach.

For several years, I then turned to Crevier (1993) for a history of Artificial Intelligence, and to selections from Bechtel & Abrahamsen (1991) for connectionism. Crevier worked adequately, but was too light and in any case is now out of print. Bechtel & Abrahamsen served well, but as connectionism has waned in its frontier importance, the detail

(continued on page 3)

## Teaching CogSci (cont.)

of coverage began to become a little inappropriate, and it could not address the most recent developments in the field. Both books, of course—through no fault of their own—violated the basic interdisciplinary character of Cognitive Science.

This year I have replaced both Crevier and Bechtel & Abrahamsen with Franklin's (1995) *Artificial Minds*. Franklin offers an explicitly historical discussion, and in this sense is highly tuned to the overall design of the course. He organizes the book around three major phases of cognitive science history, individuated by the dominant and frontier approaches to representation and mind. This is strongly parallel to my own view, and so, obviously is congenial. The first phase was the classic computational or symbol manipulation phase; the second revolved around connectionism; and the third is identified by Franklin as being constituted by situated and autonomous agent approaches. I would put more emphasis on the agents than on situatedness per se, but otherwise concur with Franklin's organization. Franklin's book is partly interdisciplinary, but makes no special point of covering the range of cognitive science disciplines. The discussions, however, often turned out to be on the light side.

In the past, I have followed Crevier and Bechtel & Abrahamsen with some of my own work. In Spring 2000, I followed Franklin with Bickhard (1996), selections from Bickhard & Terveen's (1995) *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*, and some newer papers of mine (Bickhard, in press-a, in press-b). This served both to give some substance to the ranting that I had been doing up to that time in the semester, and to give at least one sample of fully contemporary work.

These readings also transition naturally into the robotics and autonomous agent focus of the last book in the class, Clark's (1997) *Being There*. Clark's book is primarily philosophical, and so does not fully carry forward the interdisciplinary theme. But Clark is sensitive to the fact that the agentive approach itself borrows from multiple disciplines, and so the book retains an awareness of the underlying interdisciplinary nature of the field. Clark also has his own focus on these issues, one that I don't fully agree with, but the book is a good read, and makes for a successful close to the class discussions, which tend to range far beyond the explicit contents of the readings. A recent book that might also serve for this most recent historical phase is Pfeifer and Scheier's (1999) *Understanding Intelligence*, but Pfeifer & Scheier is 1) massive, and 2) focused much more on design principles than is appropriate for a strictly introductory course. Nonetheless, the next time I teach the class, I will consider using excerpts from Pfeifer & Scheier.

The historical approach has been the only way I have so far found to capture the interdisciplinary character of cognitive science while also maintaining the first year introductory character of our course. It has allowed me the flexibility to put together the necessary coverage, and to replace some of that coverage as necessary and as desirable over time.

The historical approach, however, is not just a resort of necessity but has distinct advantages. Without a sense of the history of the field, it is much more difficult to understand how and why the field, and its various parts, have arrived at their current positions. Without knowing what the historical failures of computationalism were, for example, students are left to uncover for themselves the fundamental underlying issues—and such individual level historical recapitulation is seldom a successful strategy. Without understanding the previous errors that historical shifts were motivated to solve, it can be impossible to understand why those shifts occurred at all, and that includes the shift to the most recent phase.

Science is a historical process, building on previous attempts and their successes and failures. Robotics, autonomous agents, dynamic systems, and related approaches are now at the frontiers of the field not just because of historical whims, but because computationalism and connectionism were not adequate to problems that were accepted as central to the field. That, of course, is not to claim that computationalism and connectionism have been or should be abandoned, or their study curtailed. In fact, computational and connectionist contributions have been incorporated into robots and autonomous agents. However, this is a much more subtle process than merely persisting with computational and connectionist models per se, and it is that kind of subtlety that needs to be understood by students. In my view, such understanding is not an advanced deeper level that follows on lengthy study of computationalism and connectionism per se, but is essential and extremely useful to real understanding of computationalism, connectionism, and autonomous agents, even at a rudimentary, introductory, level.

Historical approaches to scientific fields are rare outside of history courses, but they offer strong advantages nevertheless. If education is more than just filling the empty buckets on top of students shoulders with facts (Popper, 1965), but has instead to do with understanding and skill at thinking, understanding the history of a field, including its errors, offers a much deeper understanding of the field per se.

---

### Cognitive Science Dressage: Teaching Interdisciplinarity to the Undisciplined

Alex Levine
Lehigh University
( alex.levine@lehigh.edu )
http://guava.phil.lehigh.edu/ alexhome.htm

Francis Bacon is credited with having called philosophy the "Queen of the Sciences." Since this dictum gives my home discipline pride of place, I've always found it appealing. But in the Spring 1999 semester, while teaching our Introduction to Cognitive Science for the first time, I came to appreciate the true meaning of the Baconian slogan. Philosophy is indeed a queen of the sciences: a drag queen, a beautiful, seductive impostor. For a philosopher, the trick to introducing neophytes to cognitive science consists in harnessing the seductive imposture of philosophy in the service of an interdisciplinary ideal.

I opted to structure my course thematically, as opposed to historically. Two prized philosophical chestnuts, the mind-body problem and the problem of other minds, are gripping enough to furnish some sort of thematic unity even in a course whose primary thrust is not philosophical. It is possible, with some artifice, to represent much of the work done within the various contributing disciplines of cognitive science as addressed toward one or the other of these problems. Toward this end, I found Paul Churchland's dated, but eminently accessible *Matter and Consciousness* (1988) a useful introduction, which I followed with readings from John Haugeland's (1997) *Mind Design II* anthology and a collection of contemporary

(continued on page 4)

## Teaching CogSci (cont.)

readings selected to counter Haugeland's engineering bias. I originally assigned Barbara von Eckhardt's *What is Cognitive Science* (1995), but to my chagrin the students could not follow it, and the text had to be abandoned.

The thematic option has advantages and disadvantages. One advantage is that the choice of unifying themes allowed me to play to my own strengths. In my experience, students know when one is trying to pretend to be what one is not. I could not pretend to be something other than a philosopher of mind with some knowledge of how my field contributes, along with other disciplines, to the grander project called cognitive science. One obvious disadvantage of my approach was that some of the contributing disciplines of cognitive science received rather cursory treatment. Philosophy, psychology, and computer science were well represented, neuroscience and linguistics less so, and anthropology hardly at all. In cognitive science, any choice of unifying themes, my own included, inevitably imposes an illusory coherence on what, in reality, is a collection of very loosely connected research projects. Something gets lost in the process.

And so, a philosopher in drag as a cognitive scientist, I presented what I thought would be a seductive sampling of relevant work. I was aware of the masquerade, though, and honesty demanded I let my students in on it too. What consoled me was my certainty that the illusion of cohesiveness was a necessary compromise. Though I could only regret the limited scope of my own disciplinary training and experience, no doubt any psychologist or neuroscientist faced with my task would labor under corresponding limitations. More important, the very idea of teaching an introduction to cognitive science for first- and second-year students is fraught with contradiction. In order to have a successful major program, we must recruit students early. Toward this end, some entry-level course is called for. However, most members of our target audience are obviously too new to their studies to have any firm disciplinary grounding. We must teach interdisciplinarity to the undisciplined.

I suspect that this predicament remains regardless of who teaches the course, or what thematic or historical unifying framework is selected. There are various ways of addressing it, but perhaps the only way to make it disappear entirely would be to return to team-teaching the course, making it a collaborative effort by experts in several different cognitive science disciplines. As pedagogically attractive as this option seems, resource limitations prevent it from being put into practice, at least at Lehigh in 2001.

So as I look forward to teaching the introduction to cognitive science again in the Spring of 2001, my thoughts have turned to improving my drag. A true team-taught course remains an impossibility, but five colleagues in other disciplines have agreed to give guest lectures. They had better not even think about backing out. Readings are also being tweaked. The new anthology edited by Rob and Denise Cummins is promising (*Minds, Brains, and Computers–The Foundations of Cognitive Science*, 1999), and I am also considering making up for my neuroscience deficit with Gazzaniga's new *Cognitive Neuroscience: A Reader* (2000). *Mind Design II* will remain, and though I'd really like to find a more up-to-date replacement for *Matter and Consciousness*, so far none of the candidates seems appropriate.

I end with a note regarding assigned work, from which, if it's well conceived, students tend to learn more than they ever could from readings alone. Whatever misgivings I may have had regarding my choice of themes and readings, I was pleased with student reaction to my assignments, two short discursive papers and a slightly longer one. Most achieved good results on their final papers, for which I required what, for first- or second-year undergraduates, must have seemed like a lot of research. To ease the burden, I prepared an extensive bibliography and a file of key source readings, to which I referred in individual student conferences about three weeks before the due date. But I also required that all students come up with at least three sources on their own. In the end, I received a surprising number of carefully researched, well thought-out essays, whose authors had achieved a gratifying appreciation for the complexities of cognitive science research. So I conclude that successful dressage should be evaluated in terms not only of the disciplinarity, but of the scholarly *discipline* imbibed by the subject. That, after all, is much of what undergraduate education is about.

### References


Bechtel, W. and Abrahamsen, A. (1991). *Connectionism and the Mind*. Blackwell.

Bickhard, M. H. (1996). Troubles with computationalism. In W. O'Donohue, R. F. Kitchener (Eds.) *The Philosophy of Psychology*. (173-183). London: Sage.

Bickhard, M. H. (1997). Review of *The Search for Mind. Minds and Machines, 7*, 125-128.

Bickhard, M. H. (in press-a). Autonomy, function, and representation. *Communication and Cognition*, special issue on Artificial Intelligence.

Bickhard, M. H. (in press-b). Motivation and emotion: An interactive process model. In R. D. Ellis, N. Newton (Eds.) *The Cauldron of Consciousness*. J. Benjamins.

Bickhard, M. H. and Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution*. Elsevier Scientific.

Churchland, P. (1988). *Matter and Consciousness*. MIT/Bradford

Clark, A. (1997). *Being There*. MIT/Bradford.

Crevier, D. (1993). *AI*. Basic Books.

Cummins, R. and Cummins, D. (1999). *Minds, Brains, and Computers–The Foundations of Cognitive Science*. Blackwell.

Franklin, S. (1995). *Artificial Minds*. MIT.

Gardner, H. (1985). *The Mind's New Science*. Basic Books.

Gazzaniga, M. (2000). *Cognitive Neuroscience: A Reader*. Blackwell.

Haugeland, J. (1997). *Mind Design II*. MIT.

Kay, E. J. (1992). Lehigh revises Cognitive Science major. *CogSci News*, 5 (1).

Malt, B. and Melchert, N. (1988). "CogS 101"—Lehigh's gateway course in Cognitive Science. *CogSci News, 1(2)*.

Ó Nualláin, Sean (1995). *The Search for Mind*. Ablex.

Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence*. MIT.

Popper, K. (1965). *Conjectures and Refutations*. New York: Harper & Row.

Stillings, N. A., Weisler, S. E., and Chase, C. (1995). *Cognitive Science: An Introduction*. Bradford/MIT.

Von Eckhardt, B. (1995). *What Is Cognitive Science?* MIT.

# Representational Structure and the Mathematical Foundations of Cognitive Science

Mike Casey

Department of Psychology
Rutgers University, Newark

An ill-understood notion of representation lies in the foundations of cognitive science. If an agent produces a given collection of behaviors, that is, has a certain functionality, must it manipulate internal symbols that represent objects and features of the world, or should its abilities be viewed as a result only of the causal organization of the system, independent of any type of symbol manipu ltion or representation? If an agent is implicitly designed through an evolutionary process, to what extent is it merely interacting with its environment rather than modeling it, and to the extent that it must model its environment, what form must these models take? We will introduce a methodology for studying these issues that is mathematically well-founded, and show how to apply this methodology to a simple example of a representationally problematic system.

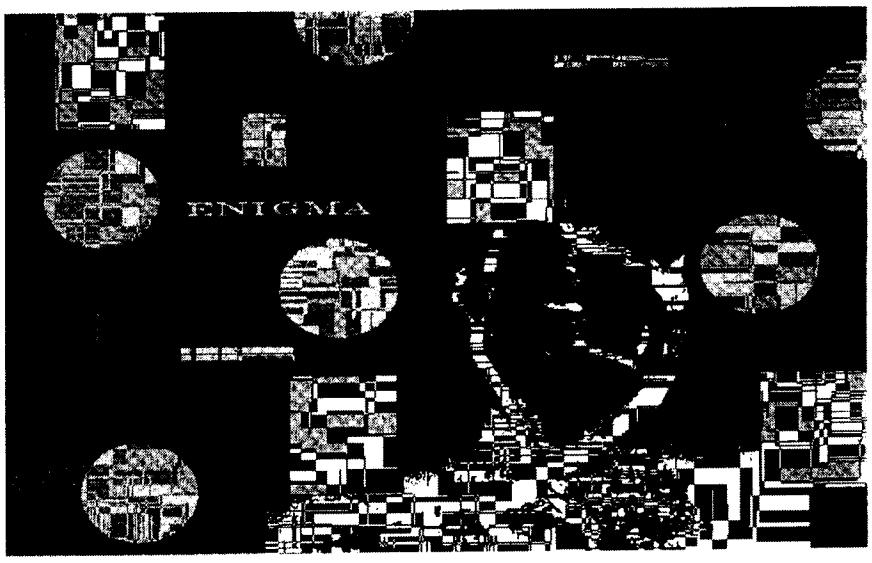
If the goal of cognitive science is to describe intelligent behavior in terms of information processing that is not dependent on the specific physical instantiation of the behavior (cf., Stillings et al., 1987), the notions of computation and representation are central. Computation provides a mechanistic, information processing framework for discussing behavior, and representation provides a potential bridge between the internal and external behavior of a system. But these notions are far from secure in their philosophical seats as the foundation of cognitive science. Other foundations for studying intelligence have been proposed, from purely behavioral (Skinner, 1957; Watson, 1930; Brooks, 1991) to the dynamical (Port and van Gelder, 1995; Beer, 1995), to the intentional (Dennett, 1987), to the neural (Churchland and Sejnowski, 1992), to the symbol-systemic (Newell and Simon, 1976). In part, this competition for a foundational language is a result of the multitude of imprecise metaphors and intuitions that have driven investigations of representation.

Rather than trying to talk about representations metaphorically, one should build theories that combine representational and computational approaches to create what we will call representational structure theories. Representational structure theories have, at their top, formal behavioral descriptions, at their bottom, abstract system descriptions, and in the middle is the minimal information theoretic structure (inferred by rigorous mathematical methods) that connect the top and bottom levels of description. The characterization of representation that points to the existence and nature of representational structure is this:

Suppose a given input/output behavior of a system is such that the form of the information in the input that is required for producing the output is not directly accessible to the output subsystem. Such a system must contain representational structure. Our description of where to look for representational structure will become more clear as we discuss the examples, but one especially important element of this characterization is that of being directly accessible. One way for a behavior to imply representation is for the output to lawfully depend on information that occurred at some point in the input past (in which case the information is inaccessible due to unidirectional nature of time's arrow). The representational structure description in this case will be a description of the system's memory and processing requirements. Another situation that would require representation would be the existence of lawful dependencies in the spatial arrangement of information that the output subsystem could not directly (i.e. without the use of intermediate variables) use to produce the output. Since the first type of inaccessibility has been studied elsewhere (Casey, 1996), the example that we give in this paper deals with the second type.

The precise form that that representational structure will take is going to depend on the constituent structure of the system, the level of analysis, and the nature of the behavior, but the following should always hold:

(Continued on page 6)




Computerized image by Christina Shepherd

## Foundations (cont.)

*Representational structure should always be the minimal element in an equivalence class of system structures that could produce the input/output behavior.*

The main intuition that tells us why representational structures should be minimal is that elements that are not minimal contain implementational details that could change without affecting the behavior of the system. We want to be careful about assigning representational power only to those elements of the system description that are essential to the system in the sense that they afford the system some functionality. The distinction between implementational details and representational structure is important for distinguishing between structure that might be of interest to someone looking inside the system and structure that is "doing representational work" for the system.

This implies that representations are not uniquely specified by vague behavioral or system descriptions. A question like "are thermostats representational?" (McCarthy, 1979) is too imprecise to be given a satisfying answer. As will be seen in the example, the precise system and behavioral description will determine the representational status of the system. In general, there will be many levels of representational structure that will be of interest, but we emphasize that this does not imply that representational structure does nothing for the system itself. This element of observer dependence in forming representational theories merely admits that one can create bad theories. If one ignores enough of the facts of the behavior, or if one attributes enough unexaminable processing power to the system (e.g. the "soul" creates reason, and to create reason you need a soul, end of story[1]), one can conclude that the system does not need causal representational structure. What one cannot do, we will show, is to agree that the system has a particular nontrivial input/output behavior and agree to a system description that is sensitive to the complexity of behavior, and then conclude (by lack of imagination or by force of ignorance) that the system contains no representations in the sense that we have defined.

Our example gives a representational analysis of the well-studied case of a feedforward neural network solving the XOR problem. This example will clarify the sense in which the hidden layer of a feedforward network forms an internal representation for the system, and show how representational structure can change as we give different conceptions of the system and the behavior. We will then explore the implications of this view of representational structure for the practice of cognitive science and will discuss mental representations, as opposed to common representations, in this framework.

### Hidden Layer Representations

The structure of a standard three layer network is commonplace enough by now that we will not give a detailed definition, but a much reduced conception that will be useful for us is the following. Let x be an input vector, z be a hidden state vector (corresponding to hidden unit activations), and y be an output vector. Further, define the functions H and O such that y = O(z) and z = H(x). That is, H is the vector valued function that transforms inputs to hidden unit activations, and O is the vector valued function that maps hidden unit activations to outputs. This is our first system conception, which we will soon refine.

What is essential to this description as a system with potential representational structure is that it contains intermediate variables of some type. That is, it contains potentially measurable states that are causally dependent on the input and causally efficacious in producing the output. What a representational theory will do is ultimately tell us something about the organization of these states implied by the input/output behavior of the system and the system description. If we would rather not remain agnostic about the actual measurements of the internal states, a representational structure theory will tell us what is important about the organization of these measured states with respect to the system's behavior. That is, it will tell us what to look for in the data and tell us when we've accounted for the behavior in question.

| Input 1 | Input 2 | Output |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

*Figure 1. The XOR problem.*

Consider our abstract feedforward neural network and the XOR problem (Figure 1).

The representational analysis begins with the question: What irreducible information configurations must exist in the system's internal variables to allow it to compute the XOR function? For the level of description given so far, the only information that they must carry is to make a distinction between 0 0, 1 1 and 0 1, 1 0. The reason for this is that the hidden unit activations act as the input to the output units, therefore, if two inputs are mapped to the same hidden unit activations, then they will also be mapped to the same outputs, but these two groups must be mapped to distinct outputs, and so must be distinct in the hidden layer. The minimal description can be depicted as show in Figure 2, which is read as follows:

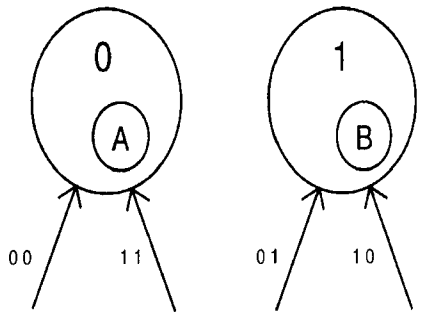



*Figure 2. A diagram of the representational structure of our first abstraction of a feedforward neural network. The circles symbolize equivalence classes of hidden unit activation vectors that produce the output indicated by the values and are activated by the inputs that label the arrows leading to the state.*

An input of 0 0 or 1 1 causes hidden unit activation equivalent to state A, that causes an output of 0, and an input of 0 1 or 1 0 causes hidden unit activation equivalent to state B, that causes an output of 1. States A and B are abstract information processing states that are instantiated in the hidden layer state vector in any number of ways, so long as they are disjoint sets and the output subsystem maps the actual states to 0 or 1 respectively.[2] The system must represent the difference between 0 0, 1 1 and 0 1, 1 0, but need not represent

(continued on page 7)

## Foundations (cont.)

the difference between 0 0 and 1 1 or 0 1 and 1 0. This means that even though there may be a different pattern of activation for, say, 0 0 and 1 1, that is not a representational distinction for the system, but rather is potentially a representation of something for an outside observer. The system only "knows" about two types of inputs.

Since we are, thus far, allowing H and O to be arbitrary functions, this is all that we can say about the information requirements of the hidden layer. If we allowed the output to be a direct function of the input, then even this simple representational structure would vanish, since then the system's internal variables would not have to make any distinctions for the system, and could even be removed. The information would be directly accessible to the output subsystem in its input form.

But the characterization given thus far is not very satisfying as a characterization of the standard feedforward network,

| Input 1 | Input 2 | Output |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |

*Figure 3. A boolean function that can be implemented without a hidden layer.*

since, for example, there is something special about XOR that makes it a more difficult computation for standard neural networks than other two-variable predicates. It is well known that a perceptron requires a hidden layer to perform the XOR mapping (Minsky and Papert, 1988). The description thus far would lead to an equally complex representational description for the boolean function shown in Figure 3.

This mapping is linearly separable, and so it can be performed by a perceptron without the use of a hidden layer representation. One might like a representational theory that distinguishes between these two behaviors. To accommodate this wish (a vaguely augmented behavioral description), we need to fill out our feedforward neural network description to include the facts that H and O can perform only linearly separable mappings. For a nice discussion of linear separability, one can consult Minsky and Papert (1988), but the intuition is that two disjoint collections of points in the hypercube form a linearly separable set if there is a hyperplane that cuts through the cube to leave all of the points in the first collection on one side and all of the points in the second collection on the other. For XOR, this means that there would have to be a line that one could draw in the unit square that left 0 0 and 1 1 on one side, and 0 1 and 1 0 on the other. It is geometrically clear that this is impossible (and can be algebraically proven to be impossible).

With this more specific system class, the hidden layer is no longer an option, but a requirement for XOR. If we allowed connections directly from the input to the output nodes, the system's representational structure would not be trivial because the information in the input is not directly accessible to the output nodes for producing the behavior. The system must transform some information contained in the input into hidden layer states so that the necessary distinction is linearly separable. Figure 4 shows an example of how this might be done. In Figure 4 we see that the system takes the initially nonlinearly separable problem and transforms it into a linearly separable problem, where the

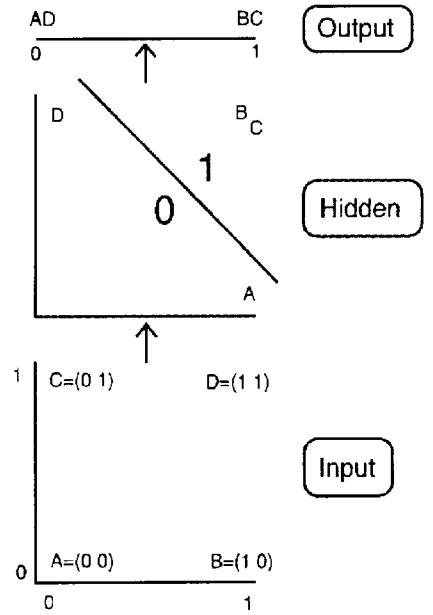



*Figure 4. Diagram of a feedforward network solution to the XOR problem. The bottom element is the input space, that only uses the corners since the function is only defined for the boolean values. The middle element is the hidden unit activation space. There are two hidden units, and the four inputs result in the four activations shown. The line, "0" and "1" show which points will be mapped to "0" and "1" outputs by the linear threshold unit output. The top element shows the resulting output.*

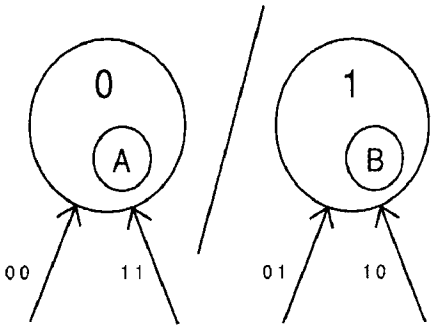



*Figure 5. A diagram of the representational structure of our refined abstraction of a feedforward neural network. In addition to the structure captured in the diagram in Figure 2, the slash between the states indicates that the vectors in these equivalence classes must be linearly separable from one another.*

mapping from the input to each hidden unit dimension is linearly separable. For example, the mapping from input to the first hidden dimension (left to right) takes D to 0, and A, B and C to 1. These input to hidden subsystems and the hidden to output subsystem conspire to solve the problem that none could solve on its own. The representational structure can be depicted as in Figure 5.

The slash mark denotes the fact that these states must be linearly separable. Otherwise, the figure is read as the previous representational structure diagram.

It would be possible to continue to refine the behavior and continue to refine the system in this way, and for some questions that one might ask about feedforward neural networks this might be necessary. For example, one could also infer that the hidden layer representation for performing the XOR function must necessarily have two distinct points for either 0 0, 1 1 or 0 1, 1 0 since if these two classes of input were

(continued on page 8)

mapped to exactly two points in hidden unit activation space then that would imply that the input to hidden unit perceptron was able to solve the XOR problem, which is impossible. One would then want to create a representational structure diagram technique that captures this indeterminacy and interdependency between representational states.

At this point, however, we do not have a nontrivial question that this inference is useful for understanding, so there is no use in making this distinction—whether 0 0, 1 1 or 0 1, 1 0 or both are represented by one or two points is an implementational detail. Also, whether these points are embedded in a one dimensional space or a one million dimensional space (one[3] or one million hidden units), is a purely implementational detail since we cannot tell without looking inside or inferring from additional information which is the case. The only question we were trying to answer with this representational structure analysis is what representational role the hidden layer plays in producing the XOR mapping, and a cluttered theory is a bad theory.

The contrast between the initial, indiscriminate and the refined neural network models shows that there is not a single representational description for a given functionality. The representational structure will depend upon the level and focus of description of the behavior and the system, but these representational descriptions will be related when they are derived from refinements of the behavioral and system descriptions. Hence, representational structures implied by a behavioral capacity are likely to have a family resemblance when the model class of the system is changed, but this family resemblance could require difficult mathematical inference to see.

This partial analysis should begin to make it clear that most neural network studies that claim to be examining hidden unit representations simply confound real representational structure and what cannot reasonably be called anything but implementational details. Just looking at a pattern of activation of hidden units can be considered a representation of the input for us, but precious little might be representational for the system. The same holds for neuroscientific investigations of "neural representations." Neural representations are possibly representational for us in some way, but it is not clear what part of the data is representational for the system.

At this point, it would be useful to embark upon a less trivial representational structure analysis to see that it is a productive approach to more complex behaviors. Fortunately, such an analysis already exists for dynamical systems (including recurrent neural networks) performing algorithmic sequential computations. Unfortunately, presenting such an analysis is beyond the scope of this paper. Fortunately, the interested reader can find the details of that theory in Casey (1996). While symbol manipulation is more commonly associated with the sequential computations studied in that work, if we take symbol manipulation to mean the formation of information carrying states that afford the system the desired behavior, then it is clear that even our feedforward network is manipulating symbols when it has nontrivial representational structure.

## Evaluation of Neuroscientific and Other Approaches to Cognitive Science

If, as we have suggested, the system class in which the behavior is to be instantiated is an important element of representational structure theories, then it might, at first glance, seem to be the case that we are advocating a full neuroscientific investigation into the nature of intelligence. But this is not the case. Neuroscience is an excellent approach to understanding medical phenomena pertaining to the nervous system, since sickness is often an expression of "details gone awry," which creates maladaptive behavior,[4] but it is less clear that intelligent behavior and mental representation depend on these details. Until the representational inference process and essential behavioral characteristics of intelligence and mental representation are better understood, it will be difficult to separate essential neuroscientific and other biological data from implementational details.

Representational structure theories can provide an amazing amount of "data reduction" in our understandings by focusing our attention on what's essential for the functionality. In the case of dynamical systems and algorithmic sequential computations (Casey, 1996), in spite of the incredible variety of potential representational structures (arbitrary geometries and topologies of state spaces, various attractor types such as chaotic and quasiperiodic, measure theoretic properties, etc.), the only structure that turned out to be essential for the system behavior was the existence of the appropriate equivalence classes of states, and the appropriate dynamical update rules for permuting them. Hence, it is easy to spend too much time studying systems before getting to the real work of understanding how the system produces functionality.

If one is truly interested in understanding intelligence rather than neuroscience, then one should conscientiously ask whether some data is likely to tell us about something that cannot easily be done some other way. Too much of neuroscience is an exercise in seeing if one can produce a trivial behavior with an overly complicated system with terrible coordinates, as is the case for the most thoroughly understood neuroethological systems (Konishi, 1991; Heiligenberg, 1991). These studies are excellent science as neuroethology, but are incredibly inefficient and wrong-headed approaches for cognitive science. This is because the behavioral yield is minimal when compared to the effort and because, thus far, neuroscience specifically focuses on many implementational details and leaves these in theories as if they were essential players.

The representational structure approach, on the other hand, focuses as much attention as possible on developing theories, rather than models, of how to produce intelligent behavior. But it can be very difficult to develop such theories. It is often easier to write a program or build a machine that can produce some intelligent behavior of choice rather than creating a general abstract theory of the behavior. This is one place where we must separate science from engineering to understand motivations. Engineering is satisfied to create some number of instantiations of a behavior, not necessarily saying anything about how nature or some other implicit designer might accomplish the task. Science, on the other hand, should strive not only to build models, but to understand how intelligence works in its greatest generality. We do not, however, mean to imply that representational structure theories are pedantic or merely decorative. In order to reach a certain level of proficiency in designing complex intelligent systems, it will be necessary to have well-developed theories of system functionality of the type that we have described. Artificial intelligence is sorely in need of a theoretical foundation and scaffolding to support its

(continued on page 9)

## Foundations (cont.)

efforts. Representational structure theoretic results will allow projects with similar behavioral goals to share knowledge by identifying the essential elements of programs and machines designed to produce those behaviors.

Another use for this approach would be to allow for a dispassionate and rigorous investigation into the idea that finding just the right neuron or just the right synapse or just the right synaptic modification rule will quickly lead to an understanding of the brain and mind. We claim that this idea is at least as absurd as the idea that knowing the correct computer programming language commands will make writing all useful programs a trivial exercise. There's little reason, from the standpoint of complexity, to think that intelligence is likely to just fall out of the details of enough biology. If we had waited for an understanding of how the brain calculated before building a calculator, we'd know a little more about the brain but we certainly wouldn't have desktop computers. One might, of course, argue that important classes of behavior require certain varieties of dynamical structures in any dynamical systems that instantiate them and then show that the neuron or synapse or circuit that is being investigated has a real chance of giving those types of structure. But until we have better ideas about representational structures, or even which behaviors are relevant to intelligence, there does not seem to be a pressing need, from the standpoint of cognitive science, to try to understand how the brain implements behaviors.

Another belief that is fueling current neuroscientific philosophizing is that neuroscience offers an alternative view to the symbol-system metaphor of the mind. There are very good reasons to think that we have to spend more time studying classes of systems besides symbol-systems, but this doesn't mean that we need to look to the nervous system. There is no real shortage of abstract systems that deserve further investigation and further attempts at integration into cognitive science. Physics, as a mature science, allows for a wide variety of system classes in its models. Considering the vast complexity and variety of intelligence, there is little reason to suspect that only symbol-systems, or only differential equations, or only neurobiological systems will suffice for all understandings of intelligence.

### Mental Representations

How, therefore, should one study mental representations? We take mental representations to mean representation that is accompanied by awareness. Mental representations, we conjecture, are a special class of the mundane representations contained in the representational structure framework that we have been studying. What distinguishes mental representations is that they are implied by a certain type of information processing, and by certain classes of system instantiations that include information processing structures that can further be inferred to imply awareness. That certain styles of information processing might imply mental representations is clear to anyone who believes that they can tell whether someone is truly paying attention to them or not. It should be clear to anyone who has ever attributed consciousness to a program or some other simple automaton that there is a great need for more precision in this area to decide when this inference is valid.

The ability to infer the existence of awareness is not a new idea. When we try to tell whether or not a person is lying, we are playing this inference game. What is new, based on our theory, is the framework for approaching this inference. Our framework advocates strong psychophysical and other behavioral research as especially relevant for understanding mental representations. The goal of these behavioral investigations must be to give an essentially abstract characterization of what humans can and can not do under aware and unaware conditions. In addition, we have the less widely accepted implication that it is of crucial importance to formulate mathematically precise information processing abstractions of these behavioral phenomena in order to keep ourselves honest about the inference process and that it is also crucial to create abstract causal system characterizations of these tasks. We conjecture that the awareness part of mental representations will be describable in terms of representational structure (or possibly a meta representational structure) for some abstract class of behaviors and some abstract class of system structures, and that neither piece alone will give a viable theory of awareness.

Why must the relevant behavioral descriptions be essentially abstract? One thing that suggests this is that expertise in a task often makes relevant information processing go from being conscious to unconscious. For example, if we take Penfield's automaton[5] (Penfield 1975:39) seriously, then we might conclude that nothing about driving a car, except maybe obeying traffic signals, implies conscious awareness. But when learning to drive a car, it seems that there is a great deal of conscious processing that must take place. So the behavioral phenomena that imply awareness must, at the very least, capture what is different about driving a car when it is an unfamiliar task versus when it is a familiar task. Another means of narrowing the focus of awareness implying processing is to look to pathologies of awareness, such as hysterical blindness, to see what kind of processing can be inferred to exist in spite of a lack of awareness of these abilities. Hence, much of the information processing behavior that is typically accompanied by awareness will be of no use for inferring the existence of awareness. Something about the range of uses for a given piece of knowledge seems to be an intuitively satisfying first approximation to the distinction between awareness implying, and awareness neutral. This bodes well for representational structure theories as a framework for studying awareness since it means that the type of processing is an essential element of awareness, and representational structure theories explicitly include processing structure.



Why do we explicitly include the possibility that the awareness of mental representations is dependent on the instantiation? The abstract behavioral description suggests that it might be something about the way information is fed back into and shared within the system

(continued on page 10)

## Foundations (cont.)

efforts. Representational structure theoretic results will allow projects with similar behavioral goals to share knowledge by identifying the essential elements of programs and machines designed to produce those behaviors.

Another use for this approach would be to allow for a dispassionate and rigorous investigation into the idea that finding just the right neuron or just the right synapse or just the right synaptic modification rule will quickly lead to an understanding of the brain and mind. We claim that this idea is at least as absurd as the idea that knowing the correct computer programming language commands will make writing all useful programs a trivial exercise. There's little reason, from the standpoint of complexity, to think that intelligence is likely to just fall out of the details of enough biology. If we had waited for an understanding of how the brain calculated before building a calculator, we'd know a little more about the brain but we certainly wouldn't have desktop computers. One might, of course, argue that important classes of behavior require certain varieties of dynamical structures in any dynamical systems that instantiate them and then show that the neuron or synapse or circuit that is being investigated has a real chance of giving those types of structure. But until we have better ideas about representational structures, or even which behaviors are relevant to intelligence, there does not seem to be a pressing need, from the standpoint of cognitive science, to try to understand how the brain implements behaviors.

Another belief that is fueling current neuroscientific philosophizing is that neuroscience offers an alternative view to the symbol-system metaphor of the mind. There are very good reasons to think that we have to spend more time studying classes of systems besides symbol-systems, but this doesn't mean that we need to look to the nervous system. There is no real shortage of abstract systems that deserve further investigation and further attempts at integration into cognitive science. Physics, as a mature science, allows for a wide variety of system classes in its models. Considering the vast complexity and variety of intelligence, there is little reason to suspect that only symbol-systems, or only differential equations, or only neurobiological systems will suffice for all understandings of intelligence.

### Mental Representations

How, therefore, should one study mental representations? We take mental representations to mean representation that is accompanied by awareness. Mental representations, we conjecture, are a special class of the mundane representations contained in the representational structure framework that we have been studying. What distinguishes mental representations is that they are implied by a certain type of information processing, and by certain classes of system instantiations that include information processing structures that can further be inferred to imply awareness. That certain styles of information processing might imply mental representations is clear to anyone who believes that they can tell whether someone is truly paying attention to them or not. It should be clear to anyone who has ever attributed consciousness to a program or some other simple automaton that there is a great need for more precision in this area to decide when this inference is valid.

The ability to infer the existence of awareness is not a new idea. When we try to tell whether or not a person is lying, we are playing this inference game. What is new, based on our theory, is the framework for approaching this inference. Our framework advocates strong psychophysical and other behavioral research as especially relevant for understanding mental representations. The goal of these behavioral investigations must be to give an essentially abstract characterization of what humans can and can not do under aware and unaware conditions. In addition, we have the less widely accepted implication that it is of crucial importance to formulate mathematically precise information processing abstractions of these behavioral phenomena in order to keep ourselves honest about the inference process and that it is also crucial to create abstract causal system characterizations of these tasks. We conjecture that the awareness part of mental representations will be describable in terms of representational structure (or possibly a meta representational structure) for some abstract class of behaviors and some abstract class of system structures, and that neither piece alone will give a viable theory of awareness.

Why must the relevant behavioral descriptions be essentially abstract? One thing that suggests this is that expertise in a task often makes relevant information processing go from being conscious to unconscious. For example, if we take Penfield's automaton[5] (Penfield 1975:39) seriously, then we might conclude that nothing about driving a car, except maybe obeying traffic signals, implies conscious awareness. But when learning to drive a car, it seems that there is a great deal of conscious processing that must take place. So the behavioral phenomena that imply awareness must, at the very least, capture what is different about driving a car when it is an unfamiliar task versus when it is a familiar task. Another means of narrowing the focus of awareness implying processing is to look to pathologies of awareness, such as hysterical blindness, to see what kind of processing can be inferred to exist in spite of a lack of awareness of these abilities. Hence, much of the information processing behavior that is typically accompanied by awareness will be of no use for inferring the existence of awareness. Something about the range of uses for a given piece of knowledge seems to be an intuitively satisfying first approximation to the distinction between awareness implying, and awareness neutral. This bodes well for representational structure theories as a framework for studying awareness since it means that the type of processing is an essential element of awareness, and representational structure theories explicitly include processing structure.



Why do we explicitly include the possibility that the awareness of mental representations is dependent on the instantiation? The abstract behavioral description suggests that it might be something about the way information is fed back into and shared within the system

(continued on page 10)

that creates awareness. With different system architectures it should be possible to modify this feedback to avoid consciousness. For example, one might make the elements much faster or numerous and then arrange them to overcome whatever architectural feedback arrangements produce awareness. Conversely, a system might achieve mental representations based on fewer behavioral constraints if the right architecture were chosen. Hence, awareness may be the byproduct of fitting certain ambitious information processing abilities into a system with resource limitations of various types.

As a further benefit to the study of mental representations, we may be able to turn this approach to mental representations on its head and use this line of reasoning to understand why mental representations are the way that they are. Why would the qualia associated with the color red be different from those of the sound of a trumpet, or, even better, an orchestra? If, as representations, these qualia are minimal information carrying structures made accessible to some system,[6] then the information in the qualia should be consistent with the desired behavioral affordance of the qualia. One characteristic of awareness, familiar since at least the time of Hume, is that anything that can be made conscious can be associated with anything else that can be made conscious. So, minimally, mental representations should be limited to those things that can be safely associated with one another. A pure blue sky should not afford as many associational opportunities to the average perceiver as the perception of the actions and qualities of a potential mate. Aberrations in the informational complexity of qualia are a trademark of mental illness, e.g., schizophrenia (too many) or denial (too few). Further, that information that should be allowed to be flexibly associated with other aspects of an agent's representational structures (e.g. informational abstractions of situations that are threatening to one's self) should be mentally represented (hence the feeling of fear in addition to reflexive fear conditioning).

The processing structure of the functionally defined "mental representation" system should largely account for the behavior and system structure that implies awareness. Hence, we propose that a solution to the "hard problem" (Chalmers, 1996) should look to representational structure in the proper range of system classes (that include the right types of feedback) for the abstract information processing acts that are always associated with awareness. A strong indication that representational structure theories are good candidates for awareness implying theories is that if one accepts that awareness is a product of the brain, then many non-identical brains produce awareness, and hence a real explanation of awareness should point to the essential commonalities among brains. This level of essential commonalities based on functional implications and structural constraints is precisely the domain of representational structural theories.

## Conclusion

We have shown that the idea of representation can be given a firm mathematical foundation and that it fills a non-vacuous role in our understanding of cognitive systems having many interesting levels of input and output and complicated interactions between them. For some classes of behavior we expect the notion of representation to play a central role, while in others the need for representational structure theories will be less pressing. Some classes of control systems, for example, and many explicitly engineered systems[7] will not be well served by representational theories. But for something like language, which has resisted much computational characterization, moving into a wider class of systems with real representational theory in mind should lead to important insights.

Cognitive science is not equally concerned with all representational systems. For studying intelligence, one should focus on refining the classes of behaviors that define intelligence, and create representational structure theories for these behaviors in broad, mathematically natural, classes of systems. By choosing broad classes of systems, one has the ability to study functionality in a way that shares the understanding between the widest possible range of implemetations (which is a stated goal of cognitive science). While this approach makes no explicit claims about how to characterize intelligence, an implication that this approach has for the Turing Test is that a very interesting line of research would be to ask which would be the best types of questions to ask in order to ascertain whether they are talking to an automaton of relatively simple design or an intelligent individual. Questions that stress complex mental imagery that also require the ability to do on the fly abstraction, for example, would probably be excellent. But even with the best set of questions, all we will be able to infer from the queries and responses alone is a certain confidence that the queried agent is conscious or intelligent in a way that is similar to a typical (or even exceptional) human. Hence, rather than devoting our efforts to the querying process, it would be of greater scientific interest to find the collection of abstract behavioral potentials and abilities that fill out the vague notion that we call intelligence.

We propose that to study cognition, one should study representations, and that to study representations, one should develop representational structure theories. This involves giving an abstract characterization of the system and behavior, and with these in hand, determining the informational dependencies between the given input and output to infer the representational structure theory relevant to the questions of interest. This leaves one with essential causal structure responsible for the functionality, which ignores implementational details and explains how systems work even before we uncover their details: because we know that if a system has a given functionality, then the system works precisely because it instantiates the given representational structure, and because it instantiates the representational structure, the system produces the functionality.

We have shown how this could work in one case, but there will need to be a rich variety of such theories to gain a complete appreciation of the representational properties of human beings. We have given a clear division of labor among "experimental" and "theoretical" cognitive scientists (which is characteristic of mature scientific fields). Experimental cognitive scientists find ways to explore and refine intelligent functionality, while theoretical cognitive scientists develop the mathematical tools necessary for creating representational structure theories. Through providing an understanding of what is and is not justified when connecting and separating levels of description, we hope that this work serves to clarify some of the philosophical and methodological issues concerning the practice of cognitive science.

## Acknowledgments

I would first like to apologize for what must be numerous oversights in citation. The ideas in this paper seem to have come

(continued on page 11)

from further consideration and interpretation of my thesis work, prompted by good debate. But I may have seen critical insights at times when I was not able to appreciate them as such and subsequently "rediscovered" them. I have not yet had the opportunity to find the extent to which my ideas overlap with others' excepting those that have been cited.

I would like to thank Arnold Mandell for his guidance, ideas and direction that eventually led to this work. I thank Omar Haneef, Mark Bickhard, Steve Hanson, Ben Martin Bly, Gordon Bearn, and Maja Matari'c for valuable discussions. I would also like to thank Gary Cottrell, David Zipser, and Terry Sejnowski for their generosity over the years. I thank the participants in the Dynamics of Cognition Symposium held at the Santa Fe Institute July 1996 for sharing their ideas on representation, and especially thank Hillel Chiel and Randy Beer for their inspiring philosophical opposition. If truisms are not challenged in insightful ways, then they have no chance to evolve. I would like to thank Jim Crutchfield for his perspective on model classes and friendship during my visit to Santa Fe. I thank Eve Marder and Larry Abbott for allowing me the time to pursue this work at Brandeis University, and the Sloan Foundation for support through a postdoctoral fellowship in theoretical neuroscience, and Steve Hanson and Rutgers University for the instructorship I held while writing this essay.

## Endnotes

1. A less transcendent example will be given a little later.

2. Note that this shows that neural network representations in this case are not vectors, as is often assumed. The representational elements are equivalence classes of vectors (of no particular dimension), and even this generalization will be too restrictive for other behavioral classes.

3. It is also easy to show that XOR cannot be solved with less than two hidden units. We start our hypothetical range with one just to show that until we do the inference, it's easy to assume the impossible. The Chinese Room argument (Searle, 1980) and Zombie thought experiments (Kirk, 1974; Chalmers, 1996) are others examples that are almost certainly assuming the impossible, but because they have not been precise enough about the problems to allow the possibility of inference (and even if they were more precise the inference would likely be yet too difficult), they are allowed their conclusion in as much as they are assumed valid until proven invalid. (It's entirely controversial whether or not a Chinese Room could be built and whether or not the translation task should be considered to imply consciousness. The Zombie thought experiment almost certainly contains contradictions.)

4. But even here, a very abstract level of understanding may be useful or even required for treatment.

5. Wilder Penfield described a rather common condition in temporal lobe epileptics which he called automatism. These patients would suffer a seizure and lose consciousness, but still continue any habitual behavior in which they might be engaged (walking, playing the piano, etc.). One such patient, Patient C, was reported to drive (sometimes through red lights) in such a state.

6. When we say system, we don't necessarily mean a spatially localized module. Spatial localization is only indirectly related to being a system.

7. It should be obvious that it's not a pressing question to ask for a theory that describes what all dynamical systems that do what desktop computers do must have in common, since it will be very difficult to decide on a description of a generic desktop computer (unless we just decide on the very bad characterization as a big finite state machine),then attempt the very difficult process of inferring the relevant representational structures and then use these to infer how these can be implemented in the various abstract machines that characterize desktop computers well enough to capture the differences between the different types.

## References

Beer, R. (1997) The Dynamics of Adaptive Behavior: A Research Program. *Robotics and Autonomous Systems* 20:257-289.

Brooks, R. (1991) Intelligence Without Representation. *Artificial Intelligence* 47:139-160.

Casey, M. (1996) The Dynamics of Discrete-Time Computation, With Application to Recurrent Neural Networks and Finite State Machine Extraction. *Neural Computation* 8:1135-1178.

Chalmers, D.J. (1996) *The Conscious Mind : In Search of a Fundamental Theory*. New York: Oxford University Press.

Churchland, P.S. and Sejnowski, T.J. (1992) *The Computational Brain*. Cambridge, MA: The MIT Press.

Dennett, D.C. (1987) *The Intentional Stance*. Cambridge, MA: The MIT Press.

Heiligenberg, W. (1991) *Neural Nets in Electric Fish*. Cambridge, MA: The MIT Press.

Kirk, R. (1974) Sentience and Behavior. *Mind* 83:60.

Konishi, M. (1991) Deciphering the Brain's Codes. *Neural Computation* 3:1-18.

**1 + 1 = 3**

*... for large values of 1.*

McCarthy, J. (1979) Ascribing Mental Qualities to Machines. In Martin Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence*, pp. 161-195. Atlantic Highlands, NJ: Humanities Press.

Minsky, M. and Papert, S. (1988) *Perceptrons: An Introduction to Computational Geometry (Expanded Edition)*. Cambridge, MA: The MIT Press.

Newell, A. and Simon, H.A. (1976) Computer Science as Empirical Enquiry: Symbols and Search. *Communications of the Association for Computing Machinery* 19:113-126.

Penfield, W. (1975) *The Mysteries of the Mind*. Princeton, NJ: Princeton University Press.

Port, R. and van Gelder, T., eds. (1995) *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: The MIT Press.

Searle, J.R. (1980) Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3: 417-458.

Skinner, B.F. (1957) *Verbal Behavior*. New York: Appleton-Century-Crofts.

Stillings, N., Feinstein, M., Garfield, J., Rissland, E., Rosenbaum, D., Weisler, S. and Baker-Ward, L. (1987) *Cognitive Science: An Introduction*. Cambridge, MA: The MIT Press.

Watson, J. (1930) *Behaviorism*. Chicago: University of Chicago Press.

# Lehigh Events

## WORKSHOP

The invitational **Workshop on Adaptive Agent Dynamics and Cognition** was held at Lehigh University from April 23-25, 1999. The workshop was organized by Mark Bickhard (Lehigh) and attended by Randall Beer (Case Western Reserve), Peter Cariani (Harvard), Wayne Christensen (University of Newcastle, Australia), Cliff Hooker (University of Newcastle, Australia), Alvaro Romero (University of the Basque Country, San Sebastian, Spain), and Jean Toulouse (Lehigh).

## GRANT PROPOSAL

Congratulations to **Mark Bickhard and the Complex Systems Group** on the recent funding of their NSF Biocomplexity Incubation Proposal, "Complex Systems from Physics to Biology."

## COLLOQUIA: 1999-2000

22 January 1999
"Searching the Web: It's Worse Than You Thought!"
C. Lee Giles
NEC Research Institute

The World Wide Web is a revolution in information dissemination, storage, and access. It has opened up new possibilities in areas such as general and scientific information dissemination and retrieval, commerce and business, education, government, religion, law, entertainment, and health care. There are many avenues for improvement of the Web, for example in the areas of locating and organizing information. I discuss the effectiveness of Web search engines, including results that show that the major Web search engines cover only a fraction of the "publicly indexable Web." Current research into improved searching of the Web is discussed, including new techniques for ranking the relevance of results, and new techniques in metasearch that can improve the efficiency and effectiveness of Web search. The creation of digital libraries incorporating autonomous citation indexing is discussed for improved access to scientific information on the Web. (This talk described joint work with Steve Lawrence and Kurt Bollacker).

11 February 1999
"The Dynamics of Working Memory in Human Cognition"
Jack Gelfand
Princeton University

I present a model of the brain mechanisms operating in working memory tasks that is consistent with the anatomy and physiology of cortical and associated subcortical structures. This includes dynamical processes in thalamocortical loops which generate short-term persistent responses in prefrontal cortex corresponding to working memory function. The properties of computer simulations based upon this model are compared with human behavior in various cognitive tasks. The mechanism of processing in semantic recall tasks is discussed in terms of temporal synchrony of firing, binding and the problem of multiple instantiation.

19 February 1999
"Children's Teleological Reasoning"
Deb Kelemen
Pennsylvania State University

Teleological reasoning–reasoning based on the assumption of purpose, design or function–is a fundamental aspect of adult thought. It leads us to wonder about the goals underlying people's actions and to view artifacts (such as clocks) and biological structures (such as eyes) as designed for purposes. However, while teleological assumptions are a substantial feature of adult thought, there is still relatively little known about children's teleological intuitions. This talk focuses on one aspect of early teleological understanding–the scope of children's tendency to view objects as existing to perform functions. It explores the predictions of two differing points of view: One proposal argues that the tendency to view objects as "designed for certain purposes" is an innate, basic, mode of construal that from early on is selectively applied to biological traits and artifacts. An alternative view–"Promiscuous Teleology"–argues that intuitions about purpose derive form children's early understanding of agency, leading them to attribute function more broadly than adults. Several studies are discussed which support the notion that preschool and elementary children are promiscuously teleological. The implications of these findings for Intuitive Biology and Folk Psychology are discussed.

(continued on page13)

## Thanks for Your Support!

*CogSci News* is distributed free-of-charge, and we hope to keep it that way. Some recipients, however, have made **voluntary contributions** in support of the newsletter. All such help is greatly appreciated!

If you would like to make a contribution, please make your check payable to "Cognitive Science Program" and mail it to The Editor, *CogSci News*, Lehigh University, 17 Memorial Drive East, Bethlehem, PA 18015.

## Lehigh Events (cont.)

26 February 1999
"Thought as Language: A Metaphor Too Far"
Jay L. Garfield
Smith College

Natural language has often served as a metaphor for thought and as an epistemological entree to thought. This has been useful, theoretically fruitful, and even necessary. But despite their necessity to science, metaphors are always dangerous: It is tempting to treat them too literally, and to impute spurious features of the metaphor to that which it is meant to explain. In cognitive science thought has often been taken literally to be conducted in an inner language. I examine the sources and consequences of this over-extension of a useful theoretical device and consider a more careful circumscription of the use of overt language in the understanding of thought.

**7th Annual Keynote Lecture:**
8 April 1999
"Miswanting: Some Problems in the Psychology of Happiness"
Daniel Gilbert
Harvard University

One of psychology's most fundamental assumptions is that people strive to achieve subjectively pleasurable states, and that this requires that they do two things well. First, they must be able to predict how they will feel in a variety of possible futures, and second, they must act to bring about the most desirable of these futures. Although we tend to think of unhappiness as that which happens to us when we fail at the second of these tasks, much of our unhappiness is, in fact, due to our failure at the first. In other words, even when we know just how to get what we want, we don't always know how to want what we will like. Why is it so hard to want well? Why can't we look into the emotional future and know just how we will feel if certain events unfold? Why are we sometimes disappointed when we achieve our wildest dreams and delighted when we realize our most dreaded fears? This lecture describes scientific psychology's answers to these questions.

11 November 1999
"Emotional Qualia: A Phenomenology of Euphoria"
Natika Newton
Nassau County Community College

This talk examines a relatively neglected area of consciousness: emotional qualia. Two issues are explored. First, it is argued that emotional qualia are more complex than colors and tastes, and are therefore more open to analysis. Specifically, emotional states are experienced in terms of representations of goal-directed actions. Second, an apparent counterexample is examined: so-called "consummatory" pleasures, or the seemingly "passive" euphoric states produced by opioids. The counterexamples are only apparent: euphoric states are experienced by means of goal-directed action-imagery, analyzable in such a way that most aspects of emotional experience can be mapped onto specific brain mechanisms.

17 February 2000
"Speaking-about, Speaking-for, and Speaking-with"
Herbert H. Clark
Stanford University

Language has evolved over the ages for use in face-to-face conversation. Still, most models of speaking are not capable of dealing with the language of conversation. I distinguish among three perspectives on speaking: (1) speaking-about, or designing utterances to express things; (3) speaking-for, or designing utterances to express things for addressees; (3) and speaking-with, or designing utterances to express things in collaboration with addressees. The language of conversation, I argue, is speaking-with, and it cannot be reduced to speaking-about or even speaking-for. I illustrate with a range of findings on spontaneous language use.

**8th Annual Keynote Lecture:**
30 March 2000
"Invention, Insight, and Compression"
Mark Turner
University of Maryland

Human beings are inventive and creative. They achieve global insight into complicated matters. They compress diffuse arrays of knowledge into tractable and tightly integrated conceptual packets. What makes these performances possible is the basic cognitive operation known as conceptual integration, or blending. I will present structural and dynamic principles of blending and a range of examples in the arts, the sciences, and language.

2 May 2000
"Towards an Action-Based Conception of Truth"
Richard Campbell
Australian National University

Throughout the 20th Century, almost all philosophers assumed, without reflection, that the only kind of item that could be true or false is a proposition, statement, sentence, belief, or some such linguistically-structured item. This is despite the fact that people continued to speak meaningfully of true friends, true love, and being true to one's spouse, one's ideals, or oneself. This talk canvases some of the inadequacies of the 'linguistic' conception of truth, and outlines a case for recovering the primary use of the word "truth," based in action. Along the way, it points out how the prevailing 'theoreticist' outlook of philosophers is based on an inadequate grasp of the natural sciences, and how realizing that action is not just a phenomenon of human mentality can provide grounds for recognizing action as a basic metaphysical category. If 'being true to' is the basic concept of truth, then an analysis of the broader concept of faithfulness can throw light on how to understand the truth of statements in terms of the success of assertive speech-acts in attaining their intrinsic objectives.

13 October 2000
"A Glimpse, a Glance, and a Peek at Language Processing in Children"
John C. Trueswell
University of Pennsylvania

Most developmental studies of language use have focused on the ultimate interpretation that children assign sentences and phrases, resulting in somewhat 'static' pictures of children's emerging knowledge of a language. Studies of the dynamic processes underlying language comprehension are much rarer, owing in part to the lack of laboratory techniques suitable for use with children in the preschool and early school years. In this talk, I will present some recent work from my lab which examines the moment-by-moment interpretation decisions of young children (as

(continued on page 14)

## Lehigh Events (cont.)

young as age 4) by recording their eye-gazes as they visually interrogate a referential scene in response to spoken instructions. These studies reveal some striking developmental differences in processing ability, with the youngest children showing a reduced ability to coordinate relevant properties of the referential scene with linguistic properties gleaned from the input. I hope to touch upon these issues in two domains: one focusing on the ability to resolve local structural ambiguity, and the other focusing on the interpretation of pronouns.

27 October 2000
"Emotion Terms and Facial Expressions: A Cross-Cultural Comparison"
James S. Boster
University of Connecticut

This talk compares the mapping of emotion terms to facial expressions by speakers of two languages: English and Shuar, an Amerindian language spoken on the rim of the Amazon basin in Ecuador. The results provide evidence of both strong commonalties and important cultural differences in the identification of emotions in the facial expressions. The mapping of terms to facial expressions in both languages produces circumplex structure, of the sort suggested by Russell (1980). Often, emotion terms that are translations of each other can be inferred by shared reference to the facial expressions. Shuar speakers more often than American English speakers identify the photographs with descriptors that are not strictly speaking emotion terms: descriptions of emotion behaviors ('She's smiling.', 'He's laughing.', 'She is gritting her teeth.'); descriptions of other actions ('He's saying "who is there?"', 'He's making a joke.'); descriptions of beliefs or thoughts ('He's thinking of his lover.', 'She's thinking she will never get married.'); descriptions of emotionally charged events ('Her mother died.'); or combinations of the above ('He's feeling sad and smiling.'). The pattern is reminiscent of Lutz's (1987) description of the Ifaluk emotion system: facial expressions are often interpreted in terms of publicly observable interactions and behaviors rather than in terms of internal psychic states.

*If lawyers are disbarred and clergymen defrocked, doesn't it follow that electricians can be delighted, musicians denoted, cowboys deranged, models deposed, tree surgeons debarked, and dry cleaners depressed?*

## Check it out!

Lehigh's Cognitive Science Web Page ...

http://www.lehigh.edu/~incog/cogsci.html

*CogSci News*
*Lehigh University*
*17 Memorial Drive East*
*Bethlehem, PA. 18015-3068*
*U. S. A.*

*Non-Profit Organization*
***U.S. Postage PAID***
*Bethlehem, Pa.*
*Permit #230*